NOISE-RESISTANT UTTERANCE DETECTOR

FIELD OF INVENTION

[0001] This invention relates to noise-resistant utterance detector and more particularly to data processing for such a detector.

BACKGROUND OF INVENTION

[0002] Typical speech recognizers require at the input thereof an utterance detector 11 to indicate where to start and to stop the recognition of the incoming speech stream. See Figure 1. Most utterance detectors use signal energy as the basic speech indicator.

[0003] In applications such as hands-free speech recognition in a car driven on a highway, the signal-to-noise ratio is typically around 0 dB. That means that the energy of the noise is about the same as that of the signal. Obviously, while speech energy gives good results for clean to moderately noisy speech, it is not adequate for reliable detection under such a noisy condition.

SUMMARY OF INVENTION

[0004] In accordance with one embodiment of the present invention a solution for performing endpoint detection of speech signals in the presence of background noise includes noise adaptive spectral extraction.

[0005] In accordance with another embodiment of the present invention a solution for performing endpoint detection of speech signals in the presence of background noise includes noise adaptive spectral extraction and inverse filtering.

[0006] In accordance with another preferred embodiment of the present invention a solution for performing endpoint detection of speech signals in the presence of background noise includes noise adaptive spectral extraction and inverse filtering and spectral reshaping.

DESCRIPTION OF DRAWING

[0007] Figure 1 illustrates an utterance detector for determining speech.

[0008] Figure 2 is a block diagram of the system in accordance with a preferred embodiment of the present invention.

[0009] Figure 3 illustrates the steps for noise-adaptive spectrum extraction in accordance with one embodiment of the present invention.

[0010] Figure 4 illustrates the steps for determination of the inverse filter by use of the spectrum maxima and the inverse filtering operation.

[0011] Figure 5 is a plot of dB versus speech frame that illustrates speech/non-speech decision parameter before (original, curve A) and after (Noise-adaptive, curve B) noise adaptive process.

[0012] Figure 6 is a plot of dB versus speech frame that illustrates speech/non-speech decision parameter before (original, curve A) and after (Inverse MAX filtering, curve B) inverse filtering.

DESCRIPTION OF PREFERRED EMBODIMENTS

Frame-Level Speech Detection

Speech/non-speech Decision Parameter

[0013] In speech utterance detection, two components are identified. The first component 11 makes a speech/ non-speech decision for each incoming speech frame as illustrated in Figure 1. The decision is based on a parameter indicating the likelihood of the current frame being speech. The second component 13 makes utterance detection decision, using some sort of decision logic that describes the detection process based on the speech/non-speech parameter made by the first component and on a priori knowledge on durational constraints. Such constraints may include the minimum number of frames to declare a speech segment, and the minimum number of frames to end a speech segment. The present patent deals with the first component.

[0014] A preferred embodiment of the present invention provides speech utterance detection by noise-adaptive spectrum extraction (NASE)15, frequency-domain inverse filtering 17, and spectrum reshaping 19 before autocorrelation 21 as illustrated by the block diagram of Figure 2.

Autocorrelation Function

[0015] For resistance to noise, the periodicity, rather than energy, of the speech signal is used. Specifically, an autocorrelation function is used. The autocorrelation function used is derived from speech X(t), and is defined as:

$$R_x(\tau) = E[X(t)X(t+\tau)] \qquad (1)$$

where X(t) is the observed speech signal at time t.

[0016] Important properties of $R_x(\tau)$ include:

- If $X(t+T) = X(t)$, then

$$R_x(\tau+T) = R_x(\tau) \qquad (2)$$

which means that, for periodical signal, the autocorrelation function is also periodical. This property gives one an indicator of speech periodicity.

- If $S(t)$ and $N(t)$ are independent and both ergodic with zero mean, then for $X(t) = S(t) + N(t)$:

$$R_x(\tau) = R_s(\tau) + R_N(\tau) \tag{3}$$

Most random noise signals are not correlated, i.e. they satisfy :

$$\lim_{\tau \to \infty} R_N(\tau) = 0$$

Therefore, we have, for large τ:

$$R_X(\tau) \approx R_S(\tau) \tag{5}$$

This property says that autocorrelation function has some noise immunity.

Search for Periodicity

[0017] As speech signal typically contains periodical waveform, periodicity can be used as an indication of presence of speech. The periodicity measurement is defined as:

$$\rho = \max_{T_l}^{T_h} Rx(\tau) \tag{6}$$

T_l and T_h are pre-specified so that the period found would range from 75 HZ to 400 Hz. A larger value of p indicates a high energy level at the time index where p is found. According to the present invention it is decided that the signal is speech if ρ is larger than a threshold. The threshold is set to be larger than typical values of Rx(t) for non-speech frames.

Noise-adaptive Spectrum Extraction (NASE)

Outline

[0018] Applicants teach to use ρ as the parameter for speech/non-speech decision in an utterance detector. For adequate performance, the input to the autocorrelation function, *X(t)*, must be enhanced. Such enhancement can be achieved in the power-spectral representation of *X (t)*, using the proposed noise-adaptive pre-processing.

[0019] The input is the power spectrum of noisy speech (pds_signal[]) and the output is the power spectrum of clean speech in the same memory space. The following steps illustrated in Figure 3 are performed:

Step 1. Convert the spectrum into logarithmic domain.

Step 2. Remove high frequency components in logarithmic domain by recurrent filtering.

Step 3. Establish an estimate of noise background.

Step 4. Suppress the noise background from the signal, in linear domain.

Detailed Description

[0020] Sequence A consists of initialization stage. Sequence B consists of the main processing block to be applied to every frame of the input signal.

[0021] For sequence A, noise-adaptive processing initialization:

γ=0.5

γ_{MIN} = 0.0625

θ = 0.98

η = 0.37

α = 30

β = 0.016

frm_count =0

freq_nbr =256.

[0022] For sequence B, noise adaptive processing main section:

For i=0 *freq_nbr* **do**

log_sig = log_{10} (*pds_signal[i]*);

past_sm[i]= (1- γ) * past_sm[i] + *log_sig* * γ

tc = **if** *past_sm[i]* >*past_ns[i]***then** θ **else** η **fi**

past_ns [i] = (1-tc) * *past_sm* [i] + tc * *past_ns* [i]

diff = pds_signal [i] – α * 10 $^{past\_ns\,[i]}$

pds_refe = β * *pds_signal* [i];

pds_signal [i] = **if** (*diff* < *pds_refe*) **then** *pds_refe* **else** *diff* **fi**

end

frm_count = *frm_count* + *1*

if *frm_count*=10, THEN γ= γ_{MIN} **fi.**

SPECTRAL INVERSE FILTERING

Outline

[0023] The production of speech sounds by humans is dictated by the source/vocal tract principle. The speech signal *s(n)* is thought to be produced by the source signal *u (n)* (larynx through the vocal cords) modulated by the vocal tract filter *h(n)* which resonates at some characteristic formant frequencies. In other words, the speech spectrum *S(ω)* is the result of the multiplication (convolution in the time domain) of the excitation spectrum $U(\omega)$ by the vocal tract transfer function $H(\omega)$:

$$S(\omega) = U(\omega) \times H(\omega) \ . \tag{7}$$

[0024] For many speech applications, it is important to apply the inverse vocal tract filtering operation to perform analysis on the excitation signal *u(n)*.

[0025] Since equation 6 focuses on the periodicity in the range of the excitation signal only and not on the periodicity induced by the formant frequency, inverse filtering the speech signal to restitute a good approximation of the unmodulated speech signal improves the endpoint detection performance.

Detailed Description

[0026] Typically, the vocal tract filter is estimated using linear prediction techniques. The coefficients α_k of the auto-regressive prediction filter

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{8}$$

are computed by minimizing the mean-square error of the prediction error.

[0027] In the present application, instead of basing the inverse filtering operation on the often used Linear Prediction (LP) filter, applicants teach to perform inverse filtering operation based on normalized approximation of the envelope of the short term speech spectrum derived from the local maxima of the short term speech spectrum. The advantage is that applicants avoid computation of LP coefficients and its corresponding spectrum. Selecting local maxima in the short term spectrum is an extremely simple task, especially considering the low resolution of the short term spectrum (128 frequency points). Note that since we never operate in the time-domain to find an estimate of the

vocal tract filter, the inverse filtering in itself is performed in the log frequency domain (dB) and is implemented by simply removing (subtracting) from the original spectrum the estimated inverse filtering spectrum.

[0028] Determination of the inverse filter by use of the spectrum maxima and the inverse filtering operation is performed by the steps in Figure 4 and is as follows:

1. In the logarithmic (dB) domain, remove the mean spectral magnitude from the original speech spectrum.
2. In the mean removed short term frequency spectrum S(i), (i = 1...128), determine all the frequency position (p_j) whose magnitudes are maxima over a window centered around p_j and stretching N positions to the left and right of p_j.
3. In the list of peaks, add the first (i=1) and last (i = 128) frequency positions Their associated magnitudes are set equal to the mean of the first and last M x N magnitudes, respectively.
4. Remove the mean of the peak magnitudes from each peak magnitude.
5. If the largest resulting peak magnitude exceeds MAX_dB_DN, normalize all peaks so that the largest peaks magnitude becomes MAX_dB_DN.
6. The resulting inverse filtering H (i), (i =1...128) is defined as the maximum of the normalized peaks and 0 dB.
7. Remove the inverse filter from the original spectrum in the logarithmic domain U(i) = S (i) – H (i).

In applicant's preferred embodiment, applicants used the following parameter values: N= 3, MAX_dB_DN = 3.5 dB, and M =5.

SPECTRAL SHAPING

Outline

[0029] The spectral reshaping technique allows for the inverse filtering technique based on the envelope of the maxima to operate properly even when the first two formants in the speech signal are close together, such as in the /u/ or /ow/ sound. Indeed, in this case the formants being so close, there is no valley in the spectrum being determined between the maxima of the formant frequencies and the envelope spectrum resembles a large dome in the low frequency domain. The consequence of this is that the entire low-frequency spectrum is exceedingly inverse filtered and it is difficult to notice the voicing of the excitation in the resulting spectrum. The solution is to implement a detector at the input in the spectrum re-shaper 19 (see Figure 2) which operates on the noise-extracted speech spectrum and raises a flag when it detects two low-frequency formants close together. When this occurrence is found, a valley in the spectrum is artificially created between the peaks of the two formants, minimizing the amount of inverse filtering in the region between the two formants.

Detailed Description

[0030] First, the short term speech spectrum of the speech frame is normalized, with a mean equal to zero dB. Then, a battery of tests is performed to detect the presence of two close low-frequency formants. If we determine the following parameters,

σ_1 : The relative magnitude of the first estimated formant,

σ_2 : The relative magnitude of the second estimated formant,

λ_1 : Index in the frequency axis (1...128) of the first estimated formant,

λ_2 : Index in the frequency axis (1...128 of the first estimated formant,

a flag signaling the presence of two close low-frequency formants is raised if the following conditions are met:

1. $\sigma_1 \geq \tau_1$, $\sigma_2 \geq \tau_2$ and $(\sigma_1 - \sigma_2) \leq \tau$,

2. $\lambda_1 \geq \lambda_{min}$ and $\lambda_1 \leq \lambda_{max}$,

3. $(\lambda_2 - \lambda_1) \geq \delta_{min}$ and $(\lambda_2 - \lambda_1) \leq \delta_{max}$.

[0031] In applicant's preferred embodiment, the values of the parameters are set to be τ_1= 3.25 dB, τ_2 = 3.00 dB, τ = 1.25 dB, λ_{min} = 12, λ_{max} = 20, δ_{min} = 8 and δ_{max} =16.

Validation Experiments

Illustration of Functioning

Noise-adaptive Spectrum Extraction (NASE)

[0032] To illustrate the effectiveness of the noise –adaptive processing, the utterance "695-6250" was processed and the result is plotted in Figure 5. It clearly indicates that the noise-adaptive spectrum extraction substantially lowers the noise background. Curve A with the solid line is the original and Curve B with the dashed lines is with noise-adaptive spectrum extraction. It indicates that the noise-adaptive spectrum extraction has no impact on peak values in that it leaves speech signal intact. Typically, an 18dB improvement is achieved.

Spectral Inverse Filtering

[0033] To illustrate the effectiveness of the inverse filtering technique, the utterance "Taylor Dean" was processed and the normalized autocorrelation results are plotted in dB in Figure 6 for three scenarios: 1) unfiltered speech (original, curve A with solid line), 2) with classic LPC inverse filtering (curve B with dotted line), and 3) with inverse filtering using the proposed technique of inverting the vocal tract filter using envelope determined using the maxima of the spectrum (curve C with dashed line). It clearly indicates the following:

Inverse filtering significantly increases the autocorrelation of the voiced part of the signal. After normalization of the plot, this results in lowering the auto correlation of the noisy parts of the signal. Performing inverse filtering using the envelope determined by the well-chosen spectrum maxima does not degrade performance of the system. In the example given, it even enhances performance of the inverse of the inverse filtering. While it is visually almost impossible to discern the speech signal (between frames 120 and 140) using the original curve, the inverse filtering allows for an immediate distinction.

Spectral Shaping

[0034] Spectral reshaping only manifests itself in frames for which the detector signaled the presence of two close low frequency formants and while a visual inspection might not immediately show the advantage of spectral reshaping. Results presented in the following paragraph and Table 1 illustrates the additional gain that can be obtained by using the technique.

Utterance Detection Assessment

[0035] To evaluate the performance improvement due to the three methods, a speech database was collected in automobile environments. The signal was recorded using a hands-free microphone mounted on the visor. Five vehicles were used for recording, representing several automobile categories.

Table 1

Car	W/o preprocessing	W/NASE	W/NASE& INVFILT	W/NASE& INVFILT& SHAPING
ACCORD	34.96	3.91	1.07	1.02
B2300	33.40	3.19	0.76	0.45
CRV	26.91	2.67	1.45	1.07
Sentra	31.13	4.63	1.81	1.67
Venture	35.88	4.01	2.27	1.71
Average	32.46	3.68	1.47	1.18

[0036] Table 1 summarizes the test results. On average the first method reduces the detection errors by about an order of magnitude. The other two methods further reduce the remaining error by more than 50 percent.

[0037] The amount of additional reduction in the detection errors offered by the inverse filtering technique over the noise adaptive spectral extraction clearly illustrates the complementary of both techniques. While NASE helps minimizing the autocorrelation of the background noise by removing it, it does not help finding the voicing information within the speech signal. The inverse filtering technique, however, is able to extract the periodic voicing information from the speech signal, while it is insufficient to remove autocorrelation created by the background noise. In terms of noise characteristics, it can be stated the

NASE will operate efficiently on slowly time-varying noises with broad spectra (almost white), while inverse filtering is able to remove noises with sharp spectral characteristics (almost tones).

[0038] It should be pointed out that the remaining 1 percent of detection error can often be attributed to an external cause over which the endpoint detector has little control, such as paper friction or speaker aspiration.

[0039] While the invention has been particularly shown and described with reference to a preferred embodiment, it will be understood by those skilled in the art that various changes in form and detail may be made without departing from the spirit and scope of the invention.